Research

Ecography 31: 161–175, 2008
doi: 10.1111/j.2007.0906-7590.05203.x

Subject Editor: Miguel Araújo. Accepted 13 December 2007

Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation

Steven J. Phillips and Miroslav Dudík

S. J. Phillips (phillips@research.att.com), AT&T Labs Research, 180 Park Avenue, Florham Park, NJ 07932, USA. – M. Dudík, Computer Science Dept, Princeton Univ., 35 Olden Street, Princeton, NJ 08540, USA.

Accurate modeling of geographic distributions of species is crucial to various applications in ecology and conservation. The best performing techniques often require some parameter tuning, which may be prohibitively time-consuming to do separately for each species, or unreliable for small or biased datasets. Additionally, even with the abundance of good quality data, users interested in the application of species models need not have the statistical knowledge required for detailed tuning. In such cases, it is desirable to use ''default settings'', tuned and validated on diverse datasets. Maxent is a recently introduced modeling technique, achieving high predictive accuracy and enjoying several additional attractive properties. The performance of Maxent is influenced by a moderate number of parameters. The first contribution of this paper is the empirical tuning of these parameters. Since many datasets lack information about species absence, we present a tuning method that uses presence-only data. We evaluate our method on independently collected high-quality presence-absence data. In addition to tuning, we introduce several concepts that improve the predictive accuracy and running time of Maxent. We introduce ''hinge features'' that model more complex relationships in the training data; we describe a new logistic output format that gives an estimate of probability of presence; finally we explore ''background sampling'' strategies that cope with sample selection bias and decrease model-building time. Our evaluation, based on a diverse dataset of 226 species from 6 regions, shows: 1) default settings tuned on presence-only data achieve performance which is almost as good as if they had been tuned on the evaluation data itself; 2) hinge features substantially improve model performance; 3) logistic output improves model calibration, so that large differences in output values correspond better to large differences in suitability; 4) ''target-group'' background sampling can give much better predictive performance than random background sampling; 5) random background sampling results in a dramatic decrease in running time, with no decrease in model performance.

Predictive models of species geographic distributions are important for a variety of applications in ecology and conservation (Graham et al. 2004). For example, they have been applied to study spread of invasive species (Thuiller et al. 2005), impacts of climate change (Thomas et al. 2004), and spatial patterns of species diversity (Graham et al. 2006). The increasing availability of large quantities of occurrence data, especially from natural history museum and herbarium collections, is fuelling the active exploration of methods which do not require data on species absence, since absence data is missing in those large datasets and is expensive to collect. Presence-only modeling methods only require a set of known occurrences together with predictor variables such as topographic, climatic, edaphic, biogeographic and remotely sensed variables. Many techniques are available for this task, ranging from climatic envelopes and logistic regression, to boosted regression trees and multivariate regression splines. For some applications, different modeling methods may give very different predictions (Pearson et al. 2006, Randin et al. 2006), so it is important to develop guidelines for producing the most accurate models of species' distributions. A recent comprehensive comparison of presence-only modeling techniques (Elith et al. 2006) found that some new methods have better predictive accuracy than the established methods. A distinguishing feature of the new methods is that they can fit more complex models from smaller datasets, using explicit ''regularization'' mechanisms to prevent model complexity from increasing beyond what is supported by the empirical data. In this paper we focus on one of these new methods, called Maxent (Phillips et al. 2004, 2006), and in particular, some new extensions to the method and the tuning of its regularization mechanism to optimize predictive accuracy.

Tuning the parameters of a modeling method poses several challenges: when the number of tuned parameters is large, it may be prohibitively time-consuming to tune the method on each species separately. Furthermore, tuning is typically based on performance on a randomly selected held-out portion of a dataset. If the dataset is too small or highly biased, performance on the held-out subset need not be a good indicator of predictive performance. Therefore, it may

be desirable to limit species-specific or dataset-specific tuning, and use "default" settings. The use of default settings is justified provided that they have been validated over a wide range of species, environmental conditions, numbers of occurrences, and amounts of sample selection bias.

In Maxent, several settings affect model accuracy by determining the type and complexity of dependencies on the environment that Maxent tries to fit. The dependencies are described by simple functions derived from environmental variables, called "features". More complex features allow fitting more complex dependencies, but they may require more data. The complexity of dependencies is controlled by the choice of feature types, and by settings called "regularization parameters". These parameters prevent Maxent from matching the input data too closely, which is known as "overfitting" and has a detrimental effect on predictive performance (Hastie et al. 2001).

The contributions of this paper are two-fold. First, we explore tuning of Maxent settings with the comprehensive dataset used by Elith et al. (2006). The dataset contains presence-only data derived from collections of natural history museums and herbaria, together with presence-absence data obtained independently from rigorous surveys. We use the presence-only data to determine appropriate feature types and regularization parameters depending on the number of occurrence records. We then evaluate the tuned settings on the independent presence-absence data. The resulting configuration is used for the default settings of the Maxent software (Phillips et al. 2005) ver. 1.8.3 through the time of writing (at least ver. 3.1.0), with a minor exception noted in the Discussion.

In the second set of contributions, we introduce and examine the effects of several extensions of Maxent. Our first extension aims to capture more general dependencies on the environment and thus increase the predictive accuracy of Maxent models. Specifically, we introduce a new type of features – "hinge" features – to model arbitrary piecewise linear responses to the environmental variables. Hinge features are similar to "threshold" features (Phillips et al. 2006), which are used to model piecewise constant responses. An analogous step up from piecewise constant responses to piecewise linear responses has been helpful in the regression setting (Friedman 1991).

The second extension is a new logistic output format, which addresses the fact that the existing raw and cumulative formats can be hard to interpret. For many modeling applications, we are interested in the probability that the species is present, conditional on the environmental conditions. This probability of presence is what is estimated by the logistic format.

The third extension addresses the problem of sample selection bias. Occurrence data are frequently biased, for example towards areas easier to access such as areas near roads, towns, airports, and waterways (Reddy and Dávalos 2003). When the bias is large, presence-only models approximate the biased sampling distribution as much as they approximate the species distribution. This can be avoided by having the background sample reflect the same bias as the presence data (Zaniewski et al. 2002, Dudík et al. 2005) – specifically, the bias shared by the collectors of a particular group of species. To implement this idea, we use as background the set of occurrences for an entire group of species that may be captured or observed using the same methods (Ponder et al. 2001, Anderson 2003). We investigate the extent to which such data, called target-group background (Elith and Leathwick 2007, Phillips et al. unpubl.), can be regarded as a random sample from the sampling distribution and thus used to improve Maxent predictions.

The final extension simply reduces model building time. Rather than performing calculations involving all sites in the study area, we make use of a random subset of sites, called "background data", chosen to be large enough to sufficiently represent the distribution of environmental conditions in the study region.

Maxent modeling of distributions

Basic concepts

In maximum entropy density estimation, the true distribution of a species is represented as a probability distribution π over the set X of sites in the study area. Thus, π assigns a non-negative value to every site x and the values $\pi(x)$ sum to one. We produce a model of π, a probability distribution that respects a set of constraints derived from the occurrence data. The constraints are expressed in terms of simple functions of the environmental variables, called features. Specifically, the mean of each feature is required to be close (within some error bounds) to the empirical average over the presence sites. For example, for a feature "annual precipitation", the corresponding constraint says that the mean annual precipitation predicted by the model should be close to the average observed precipitation. Since the set of constraints typically under-specifies the model, among all probability distributions satisfying the constraints, we choose the one of maximum entropy, i.e. the most unconstrained one (Jaynes 1957).

Maximum entropy density estimation can also be explained from a decision theoretic perspective as robust Bayes estimation. Specifically, consider the scenario where the goal of the modeler is to optimize the expected log likelihood (see "Performance measures", below), and the only fact known about the true distribution π is that it satisfies a certain set of constraints. The strategy which guarantees the best performance regardless of π, also called the minimax strategy, is to choose the maximum entropy distribution subject to the given constraints (Topsøe 1979, Grünwald 2000, Grünwald and Dawid 2004).

To understand how π represents the realized distribution of the species, consider the following (idealized) sampling strategy. An observer picks a random site x from the set X of sites in the study area, and records 1 if the species is present at x, and 0 if it is absent. If we denote the response variable (presence or absence) as y, then $\pi(x)$ is the conditional probability $P(x|y=1)$, i.e. the probability of the observer being at x, given that the species is present. According to Bayes' rule,

$$P(y=1|x) = \frac{P(x|y=1)P(y=1)}{P(x)} = \pi(x)P(y=1)|X| \quad (1)$$

since according to our sampling strategy $P(x) = 1/|X|$ for all x. Here $P(y=1)$ is the overall prevalence of the species in

the study area. The quantity $P(y=1|x)$ is the probability that the species is present at the site x, which is 0 or 1 for plants, but may be between 0 and 1 for vagile organisms. Equation 1 shows that π is proportional to probability of presence. However, if we have only occurrence data, we cannot determine the species' prevalence (Phillips et al. 2006, Ward et al. 2007). Therefore, instead of estimating $P(y=1|x)$ directly, we estimate the distribution π. We emphasize that here x is a site, rather than a vector of environmental conditions. This treatment differs from more traditional statistical methods, such as logistic regression; later we will bring these two viewpoints together and present a new way of estimating probability of presence from the Maxent model (see below).

According to Section 2 of Dudík et al. (2004), the Maxent distribution belongs to the family of Gibbs distributions derived from the set of features $f_1, \ldots, f_n$. Gibbs distributions are exponential distributions parameterized by a vector of feature weights $\lambda = (\lambda_1, \ldots, \lambda_n)$ and defined by

$$q_\lambda(x) = \frac{\exp\left(\Sigma_{j=1}^{n} \lambda_j f_j(x)\right)}{Z_\lambda} \quad (2)$$

where Z_λ is a normalization constant ensuring that probabilities $q_\lambda(x)$ sum to one over the study area. Therefore, the value of the Maxent model q_λ at a site x depends only on the feature values at x, and hence only on the environmental variables at x. This means that the Maxent model, which we originally defined with strict reference to the set X of training sites, can also be ''projected'' to other sites where the same environmental variables are available. The Maxent distribution is the Gibbs distribution q_λ that maximizes a penalized log likelihood of the presence sites, namely

$$\frac{1}{m}\sum_{i=1}^{m} \ln(q_\lambda(x_i)) - \sum_{j=1}^{n} \beta_j|\lambda_j|$$

where the regularization parameter β_j is the width of the error bound for feature f_j and $x_1, \ldots, x_m$ are the presence sites. The first term, log likelihood, gets larger as we obtain a better fit to the data. This gives insight into how Maxent uses background data: the first term is larger for models that give more probability to the presence sites and less to the rest of the sites, i.e. models that best distinguish the presence sites from the background. The second term, regularization (also known as the lasso penalty; Tibshirani 1996) gets larger as the weights λ_j get larger. Larger weights λ_j typically mean that the model is more complex and is thus more likely to overfit. Maximizing the difference between log likelihood and regularization can be viewed as seeking a Gibbs distribution which fits the data well, but which is not too complex. The tradeoff is controlled by the regularization parameters.

For large sample sizes, the performance of Maxent (as measured by log likelihood of test data) converges to that of the best Gibbs distribution, as long as the presence sites are drawn independently at random according to π (Dudík et al. 2004). The theoretical analysis gives the best performance guarantees when the regularization parameters β_j are as small as possible, while keeping the true feature means (under π) within the error bounds. Thus, we have an incentive to obtain error bounds that are as tight as possible. To simplify the process of tuning parameters, we reduce the number of parameters from one per feature to one per feature class by setting

$$\beta_j = \beta\sqrt{\frac{s^2[f_j]}{m}}$$

where β is a regularization parameter that depends only on the feature class and $s^2[f_j]$ is the empirical variance of feature f_j, so $\sqrt{s^2[f_j]/m}$ is an estimate of the standard deviation of the empirical average. According to the theoretical guarantees we expect that $\beta \propto \sqrt{\log(\text{number of features})}$ will give good performance. However, the value of β that optimizes the theoretical bounds may not necessarily give the best model performance in practice. Therefore, we fine-tune the regularization parameter for each feature class separately, using empirical tuning as described below.

Environmental variables and feature classes in Maxent

Features in Maxent are derived from environmental variables of two types: continuous and categorical. Continuous variables take arbitrary real values which correspond to measured quantities such as altitude, annual precipitation, and maximum temperature. Categorical variables take only a limited number of discrete values such as soil type or vegetation type. Some categorical variables quantify the degree of some property (on a discrete scale), for example soil fertility. This type of variable is referred to as discrete ordinal. We will typically treat discrete ordinal variables as if they were continuous.

The Maxent software (Phillips et al. 2005) implements features of six classes: linear (L), quadratic (Q), product (P), threshold (T), hinge (H), and category indicator (C) features. Hinge features are introduced in this paper, while the other five classes were introduced in Phillips et al. (2006). Linear, quadratic, product, threshold, and hinge features are derived from continuous variables. Linear features are equal to continuous environmental variables, quadratic features equal their squares, and product features equal products of pairs of continuous environmental variables. Respectively, they constrain means, variances, and covariances of the respective variables to match their empirical values (Phillips et al. 2006). Category indicator features are derived from categorical variables. Specifically, if a categorical variable has k categories, it is used to derive k category indicator features. For each of the k categories, the corresponding category indicator equals 1 if the variable has the corresponding value and 0 if it has any of the remaining $k-1$ values.

Threshold and hinge features allow Maxent to model an arbitrary response of the species to an environmental variable from which they are derived. If f is a continuous variable then for any value h (called the knot), we define the threshold feature $\text{threshold}_{f,h}$ by

$$\text{threshold}_{f,h}(x) = \begin{cases} 0 & \text{if } f(x) < h \\ 1 & \text{otherwise.} \end{cases}$$

The forward hinge feature $\text{forwardhinge}_{f,h}$ is 0 if $f(x) \leq h$, then increases linearly to 1 at the maximum value of f:

$$\text{forwardhinge}_{f,h}(x) = \begin{cases} 0 & \text{if } f(x) < h \\ \dfrac{f(x) - h}{\max(f) - h} & \text{otherwise.} \end{cases}$$

In a similar way, we define a reverse hinge feature, which is 1 at the minimum value of f, drops linearly to 0 at $f(x) = h$, and is 0 afterwards. Examples of a forward hinge feature and a reverse hinge feature are shown graphically in Fig. 1. Forward and reverse hinge features are collectively referred to as hinge features, and we have coined this term to evoke the shapes seen in the figure. In the terminology of splines, threshold features are base functions of splines of order 1 (piecewise constant splines) while hinge features are base functions of splines of order 2 (piecewise linear splines).

As the number of features increases, Gibbs distributions become more complex, and may be more prone to overfitting. We therefore expect that more complex feature classes will require more regularization to yield accurate predictions. Many combinations of feature classes are possible; common combinations used are LC, LQC, HC, HQC, TC, and HQPTC. Linear features are special cases of hinge features, so it is redundant to use L and H features simultaneously.

Maxent output formats and logistic models

The primary output of Maxent is the exponential function $q_\lambda(x)$ that assigns a probability (referred to as a "raw" value) to each site used during model training. Raw values are not intuitive to work with though: in particular, it is hard to interpret "projected" values obtained by applying q_λ to environmental conditions at sites not used during model training. Raw values are also scale-dependent, in the sense that using more background data results in smaller raw values, since they must sum to one over a larger number of background points. For these reasons, raw values have generally been converted into the "cumulative" format (Phillips et al. 2006).

The cumulative format is defined in terms of omission rates predicted by the Maxent distribution q_λ. Specifically, we consider 0–1 prediction rules that threshold raw outputs at a level p. Each raw threshold p is transformed into the omission percentage c(p) predicted by q_λ for the corresponding rule, i.e.

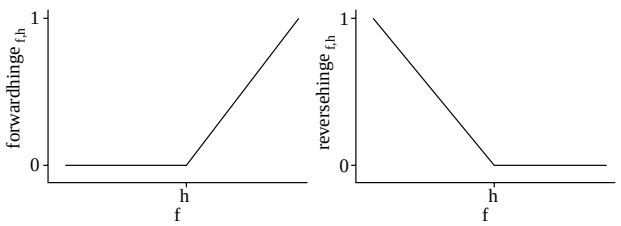

Fig. 1. A forward hinge feature (left) and a reverse hinge feature (right), defined for a continuous environmental variable f.

$$c(p) = 100 \sum_{x:q_\lambda(x) \leq p} q_\lambda(x).$$

Therefore, if we make a 0–1 prediction from the Maxent distribution q_λ using a cumulative threshold of c, the omission rate is c% for test sites drawn from q_λ. The cumulative format is scale-independent, and is more easily interpreted when projected, but it is not necessarily proportional to probability of presence.

For example, consider a generalist species whose probability of presence is close to 1 across the whole study area, with slight variations that avoid ties. Since the probability values are similar across the entire region, the cumulative values of individual sites will be roughly proportional to their rank, and hence they will range evenly from 0 to 100. Thus, big variations in cumulative value do not necessarily represent big variations in suitability or probability of presence.

We therefore introduce a new logistic output format that gives an estimate of probability of presence. Let z denote a vector of environmental variables, and let z(x) be the value of z at a site x. Traditional statistical methods such as logistic regression estimate $P(y = 1|z)$, the conditional probability of presence given the environmental conditions, which is closely related to the quantity we estimate, $P(y = 1|x)$:

$$\begin{aligned} P(y = 1|z) &= \frac{P(z|y = 1)P(y = 1)}{P(z)} && \text{(Bayes' rule),} \\ &= \frac{\Sigma_{x \in X(z)} P(x|y = 1)P(y = 1)}{\Sigma_{x \in X(z)} P(x)} \\ &= \frac{\Sigma_{x \in X(z)} P(y = 1|x)P(x)}{|X(z)|/|X|} && \text{(Bayes' rule),} \\ &= \frac{\Sigma_{x \in X(z)} P(y = 1|x)}{|X(z)|} && \text{(since } P(x) = 1/|X|) \quad (3) \end{aligned}$$

where X(z) denotes the set of locations with environmental conditions z. Therefore, in order to estimate $P(y = 1|z)$, it suffices to focus on $P(y = 1|x)$. Indeed, in the special case that $\pi(x)$ is only a function of the environmental conditions, if we let x(z) denote an arbitrary element from X(z), eq. 3 simplifies to

$$P(y = 1|z) = P(y = 1|x(z)). \quad (4)$$

Combining eq. 1 and 4, it is tempting to use our estimate q_λ of π to derive the following estimate of $P(y = 1|z)$:

$$P(y = 1|z) \approx q_\lambda(x(z))\ P(y = 1)|X|.$$

However, this approximation has two difficulties. First, we may not know or be able to estimate $P(y = 1)$, since this quantity is not determinable from presence-only data (Ward et al. 2007). Second, the approximation may result in probabilities greater than one, since Maxent does not guarantee that $q_\lambda(x)$ is smaller than $1/(P(y = 1)|X|)$.

We resolve these difficulties with a novel application of the maximum entropy principle. Rather than applying the principle to estimate a distribution over sites, we apply it to a joint distribution P(x, y) representing both a sampling distribution over sites (assumed the same for data collection and evaluation) and the presence/absence of the species. In

particular, we estimate P(x, y) by a distribution Q(x, y) of maximum entropy subject to constraints on the conditional distribution $P(x|y=1)$, i.e. the same constraints we applied to estimate π. Once we obtain the joint estimate Q, we have enough information to derive the conditional probability $Q(y=1|x)$, which turns out to be

$$Q(y=1|x) = \frac{e^H q_\lambda(x)}{1 + e^H q_\lambda(x)}$$

where q_λ is the maximum entropy estimate of π and H is the entropy of q_λ. Similarly,

$$Q(y=1|z) = \frac{e^H q_\lambda(x(z))}{1 + e^H q_\lambda(x(z))}$$

(for the derivation see Dudík and Phillips unpubl.). Thus, $Q(y=1|z)$ takes the form of a logistic regression model with the same set of parameters λ as the Maxent model and with the intercept determined by the entropy of q_λ. Because of the robust Bayes interpretation of Maxent (see "Basic concepts", above), we expect that the estimate $Q(y=1|z)$ will perform well against a range of sampling distributions and prevalence values.

The model $Q(y=1|z)$ can also be interpreted from the point of view of information theory as follows. Suppose that we receive a sequence of independent samples from the Maxent distribution q_λ, corresponding to a sequence of observations. Then the average of their log probabilities will be very close to $-H$, the negative entropy (Cover and Thomas 2006), because $-H$ is simply the mean log probability: $-H = \Sigma q_\lambda(x) \ln (q_\lambda(x))$. Thus, for "typical" sites whose log probabilities are close to this mean, we obtain $q_\lambda(x) \approx e^{-H}$. The model Q therefore assigns typical presence sites probability of presence close to 0.5.

We may have a prior expectation about the probability of presence at typical presence sites: for example, extensive collecting effort may have been required to obtain the known occurrence records for a rare species, suggesting that its probability of presence is low everywhere. This information could potentially be incorporated as a constraint on P(x, y). However, probability of presence depends on sampling effort, and in particular on site size and, for vagile organisms, on observation time. Therefore, we can more simply incorporate knowledge of sampling effort by interpreting $Q(y=1|z)$ as probability of presence under a similar level of sampling effort as was required to obtain the known occurrence data.

Note that the raw, cumulative and logistic formats are all monotonically related, so they rank sites in the same order and therefore result in identical performance, when measured using rank-based statistics such as AUC (Fielding and Bell 1997). However, their predictive performance will vary when measured by statistics that depend on actual output values such as Pearson's correlation (Zheng and Agresti 2000).

Experimental methods

Species occurrence data and environmental predictors

We used a comprehensive collection of data developed by a working group at the National Center for Ecological Analysis and Synthesis (NCEAS) as part of a large-scale comparison of species distribution modeling methods (Elith et al. 2006). We refer to the data as "the NCEAS data", and the comparison of methods as "the NCEAS comparison". The NCEAS data consists of two independent datasets for 226 species from 6 regions of the world (Table 1).

The first dataset contains presence-only data, i.e. a set of geographic coordinates of recorded presence sites for each species together with a set of environmental variables for each of the 6 regions. The environmental variables describe grids of 2 million to 27 million sites, each of which we downscaled to 1.5 million sites in order to reduce computation time. The number of presence sites per species ranges from 2 to 5822, with a median of 57. These presence-only data constitute the training data used to make models. The second dataset is presence-absence evaluation data: for each species, the evaluation data contains a set of sites of observed presence and a set of sites of observed absence. The number of test sites (presence and absence combined) ranges from 102 to 19 120. The presence-only data are derived from museum and herbarium-type collections, while the presence-absence data are derived from rigorous surveys that sample across both environmental and geographic space. For more details, see Elith et al. (2006).

Performance measures

One measure of performance applicable to Maxent is log loss. It equals the negative log likelihood of the test data, i.e. the sum of the negative log probabilities that the Maxent model assigns to test sites. Log loss is the quantity that Maxent optimizes (Dudík et al. 2004). It is always non-negative and can be arbitrarily large. Smaller values correspond to better prediction (higher likelihood). The uniform distribution over a set of N sites achieves a log loss of ln N. The true distribution π achieves the minimum log loss equal to the entropy of π, which is always smaller than

Table 1. The six study regions and their corresponding numbers of species and environmental variable types: categorical (categ.), discrete ordinal (ord.), continuous (cont.) and binary, which can be considered as discrete ordinal or categorical.

Code	Region	Taxonomic groups	Environmental variables	Species
awt	Australian wet tropics	Bird, plant	13 cont.	40
can	Ontario, Canada	Bird	10 cont., 1 categ.	20
nsw	NE New South Wales	Bird, mammal, plant, reptile	10 cont., 1 categ., 2 ord.	54
nz	New Zealand	Plant	11 cont., 2 ord.	52
sa	South America	Plant	11 cont.	30
swi	Switzerland	Plant	12 cont., 1 binary	30

or equal to ln N. We can use log loss on presence-only test data to evaluate a Maxent model, but (at least in the form used here) it is not suitable for use with presence-absence test data.

Another performance measure, applicable to any species modeling method, is the area under the ROC curve (AUC), which measures the quality of a ranking of sites (Fielding and Bell 1997). The AUC is the probability that a randomly chosen presence site will be ranked above a randomly chosen absence site. A random ranking has on average an AUC of 0.5, and a perfect ranking achieves the best possible AUC of 1.0. Models with values above 0.75 are considered potentially useful (Elith 2002). Sometimes we do not have any absence data with which to measure AUC: this is true in the present study only in our tuning experiments, where the tuning is being performed using presence-only data. In this case, an AUC can be calculated by using background data (also called pseudo-absences) chosen uniformly at random from the study area, in place of true absences. The interpretation changes accordingly: the AUC is now the probability that a randomly chosen presence site is ranked above a random background site (Phillips et al. 2006).

The correlation between a model's predictions and presence/absence in test data (regarded as a 0–1 variable) is known as the point biserial correlation, and can be calculated as a Pearson correlation coefficient (COR; Zheng and Agresti 2000). While AUC is rank-based, depending only on the relative ordering of predictions, COR measures how well calibrated the prediction values are (up to scaling and translation), in other words, whether big differences in prediction values correspond to big differences in probability of presence of the species.

Lastly, if a model is interpreted as estimating species' probability of presence, rather than just giving an index of habitat suitability, then the model predictions can be evaluated using deviance, defined as -2 times the log probability of the test data. In particular, if the model predicts a value of p for a test site, that site contributes -2 ln p to the deviance if the species is present there, and -2 ln$(1-p)$ otherwise. To combine deviance values from different regions with different amounts of test data, we normalize the deviance by dividing by the number of test sites. Like COR, deviance measures how well calibrated prediction values are, but in addition, it penalizes errors in scaling of prediction values.

Experiments for tuning Maxent

In this section, we describe the tuning of regularization parameters and selection of the best-performing sets of feature classes, using presence-only data. We selected a small subset of species in each region to use for tuning purposes. Our goal was to include a diverse set of biological groups and a wide set of sample sizes. Table 2 lists the selected species as well as experiments where they were used (experiments are described below).

In the tuning experiments, we measured the performance of Maxent for various parameter settings as follows. We randomly partitioned occurrence records of every species into a training set with 60% of the records and the test set with 40% of the records. We ran Maxent on the training set and evaluated its performance on the test set and took the average over 5–10 random partitions (see below). In all the tuning experiments, the performance is measured in terms of log loss and AUC, where the latter is measured using background data in place of true absences. Since the downscaled grids have size $N = 1.5 \times 10^6$, the uniform distribution has a log loss of ln $1.5 \times 10^6 = 14.22$. The AUC of the uniform distribution is 0.5.

Tuning regularization parameters using presence-only data (Reg)

The goal of the first set of experiments was to determine regularization parameter values for individual feature classes – called β_L, β_Q, β_P, β_T, β_C – yielding good performance in all six regions for varying numbers of occurrence records. We generated models for different feature class settings (L, LQ, LQP, T, C; hinge features are considered separately in ''Experiments for new extensions to Maxent'', below). For each setting, we assumed a single regularization parameter; in particular, for the LQ setting we kept $\beta_L = \beta_Q$ and for the LQP setting we kept $\beta_L = \beta_Q = \beta_P$.

For each feature class setting, we varied the number of occurrences (by considering nested subsets of the full training set) and the value of the regularization parameter β. The number of occurrences was chosen from the geometrically increasing sequence {6, 10, 17, 30, 55, 100, 178, 316, 1000, 3162}, and the values of β from the geometrically increasing sequence {0.02, 0.05, 0.10, 0.22, 0.46, 1.0, 2.2, 4.6} chosen to bracket the range of suitable values suggested by theory (Dudík et al. 2004). For each number of occurrences, we determined the average performance over five random partitions as a function of β. We call the corresponding plots β-curves. Peaks of β-curves (minima of log loss curves and maxima of AUC curves) correspond to optimal choices of β for each particular species. Some prior knowledge was used to restrict the set of β's and numbers of occurrences to intervals where the peaks are likely to occur; in particular, L runs were not configured on large sample sizes and LQP runs were not configured on small sample sizes. Also note that C runs were only possible for regions can and nsw, with one categorical variable in each region.

For each feature class setting and number of occurrences, we selected the best β by visual inspection of β-curves. The goal was to choose β performing well in terms of both log loss and AUC on all of the evaluated species. We first excluded curves where Maxent was not performing well: log loss β-curves where Maxent never reached performance below 14.2 (the log loss of the uniform distribution) and AUC curves that remained below 0.7 (based on the recommendations of Elith (2002) that values above 0.75 are considered potentially useful). On the remaining curves we used visual inspection to employ two strategies that could be loosely termed as ''mean criterion'' and ''median criterion''. According to the former criterion, we chose the value β that was close to the peak of β-curves of as many species as possible. According to the latter criterion, we chose β at which about half of the β-curves were increasing

Table 2. Species used for tuning regularization parameters. The fourth column gives the number of occurrence records in the presence-only data used for the parameter tuning; the last column lists sections that describe experiments in which the species was included.

Region/code	Species	Group	#PO	Experiments
awt/ausrob	*Austrochaperina robusta*	frog	193	Reg L, LQ, LQP, T; Opt
awt/bhe	*Lichenostomus frenatus*	bird	351	Opt
awt/coporn	*Cophixalus ornatus*	frog	337	Opt
awt/cryliv	*Cryptocarya lividula*	plant	44	Opt
awt/ghr	*Heteromyias albispecularis*	bird	484	Opt
awt/guiacu	*Guioa acutifolia*	plant	56	Reg L, LQ, LQP; Opt
awt/lamcog	*Lampropholis coggeri*	reptile	165	Opt
awt/sapbas	*Saproscincus basiliscus*	reptile	177	Opt
can/amcr	*Corvus brachyrhynchos*	bird	483	Opt; Cat
can/cogr	*Quiscalus quiscula*	bird	721	Reg L, LQ, LQP, C, T; Cat; Opt; Cat
can/eato	*Pipilo erythrophthalmus*	bird	119	Cat; Opt; Cat
can/gcki	*Regulus satrapa*	bird	18	Opt; Cat
can/hosp	*Passer domesticus*	bird	615	Opt; Cat
can/inbu	*Passerina cyanea*	bird	138	Reg L, LQ, LQP, C; Cat; Opt; Cat
can/modo	*Zenaida macroura*	bird	749	Cat; Opt; Cat
can/wtsp	*Zonotrichia albicollis*	bird	313	Cat; Opt; Cat
nsw/basp2	*Falsistrellus tasmaniensis*	mammal	28	Opt; Cat
nsw/dbsp2	*Calyptorhynchus lathami*	bird	426	Reg L, LQ, LQP, C, T; Cat; Opt; Cat
nsw/dbsp7	*Myzomela sanguinolenta*	bird	315	Cat; Ord; Opt; Cat
nsw/nbsp2	*Tyto tenebricosa*	bird	120	Cat; Ord; Opt; Cat
nsw/otsp7	*Eucalyptus campanulata*	plant	69	Cat; Ord; Opt; Cat
nsw/rusp2	*Cyathea leichhardtiana*	plant	42	Opt; Cat
nsw/srsp5	*Eulamprus murrayi*	reptile	186	Opt; Cat
nsw/srsp7	*Pseudechis porphyricaus*	reptile	118	Reg L, LQ, LQP, C; Cat; Opt; Cat
nz/clefor	*Clematis forsteri*	plant	36	Opt
nz/coppro	*Coprosma propinqua*	plant	40	Opt
nz/drauni	*Dracophyllum uniflorum*	plant	174	Reg L, LQ, LQP, T; Ord; Opt
nz/libbid	*Libocedrus bidwillii*	plant	105	Opt
nz/metper	*Metrosideros perforata*	plant	87	Opt
nz/metrob	*Metrosideros robusta*	plant	48	Opt
nz/phyalp	*Phyllocladus alpinus*	plant	211	Opt
nz/prutax	*Prumnopitys taxifolia*	plant	130	Reg L, LQ, LQP; Ord; Opt
sa/amphpani	*Amphilophium paniculatum*	plant	88	Reg L, LQ, LQP, T; Opt
sa/arrabrac	*Arrabidaea brachypoda*	plant	203	Reg L, LQ, LQP; Opt
sa/arracinn	*Arrabidaea cinnomomea*	plant	49	Opt
sa/cydiaequ	*Cydista aequinoctialis*	plant	138	Opt
sa/distmagn	*Distictella magnoliifolia*	plant	81	Opt
sa/fridspec	*Fridericia speciosa*	plant	57	Opt
sa/lundvirg	*Lundia virginalis*	plant	36	Opt
sa/parapyra	*Paragonia pyramidata*	plant	216	Opt
swi/abialb	*Abies alba*	plant	3357	Opt
swi/acepse	*Acer pseudoplatanus*	plant	2800	Ord; Opt
swi/betpen	*Betula pendula*	plant	468	Opt
swi/fagsyl	*Fagus sylvatica*	plant	5528	Reg L, LQ, LQP, T; Ord; Opt
swi/pincem	*Pinus cembra*	plant	279	Reg L, LQ, LQP; Ord; Opt
swi/pinunc	*Pinus uncinata*	plant	291	Opt
swi/poptre	*Populus tremula*	plant	154	Opt
swi/pruavi	*Prunus avium*	plant	613	Opt

and half were decreasing. The latter criterion was used whenever the peak was not identifiable in β-curves (they were monotone). The optimal values of β for numbers of occurrences between those in the evaluated sequence were obtained by linear interpolation.

Combining continuous and categorical variables (Cat)

In the initial block of experiments (see above), the regularization parameter for category indicators β_C was determined in Maxent runs with a single categorical variable. When categorical variables are used with additional continuous variables, the total number of features increases, so we expect that higher values of β_C will yield better performance (see the discussion in ''Maxent modeling of distributions'', above). In this set of experiments, we explored a range of alternative settings of β_C in runs including L, Q and P features derived from continuous variables.

We carried out LC, LQC and LQPC runs with β_L, β_Q, β_p equal to the previously determined optimum and for three different settings of β_C. The first β_C setting, $\beta_C = \text{low}$, corresponds to the value determined in the initial block of experiments, the second setting $\beta_C = \beta_{LQP}$ corresponds to using the same regularization for category indicators as for L, Q and P features, and the third setting, an approximation of the geometric average of the previous two settings, is an intermediate regularization choice.

For each setting of β_C, we plotted the average performance over 10 partitions as a function of an increasing number of samples from nested subsets of sizes {5, 10, 20,

40, 75, 150, 300, 750, 2000}. (This sequence differs from the sequence used in ''Tuning regularization parameters using presence-only data'', above). The rationale behind choosing a different sequence was to evaluate Maxent for the training set sizes where the interpolated values of β are used.) The resulting plots are referred to as m-curves. The best setting of β_C was again chosen by visual inspection of graphs with the goal being to perform well on all evaluated species both in terms of AUC and log loss. In the current block of experiments, we marked all discrete ordinal variables as categorical to obtain a larger number of categorical variables and hence a more reliable tuning of β_C.

Using discrete ordinal variables (Ord)

Next, we explored the effect of treating discrete ordinal variables as categorical or continuous. For the former case, we used the optimal β_C determined in the previous experiment. For the latter case, we consider two settings of β_C: the previously determined optimal setting and the baseline setting $\beta_C = \beta_{LQP}$ which uses a single regularization parameter for all features (this was the setting in versions of Maxent prior to 1.8.3). The optimal setting is determined by visual inspection of m-curves for LC, LQC and LQPC runs.

Choosing optimal feature sets (Opt)

The final goal of the tuning experiments was to decide which sets of feature classes to use for what numbers of species occurrences. We used the previously determined regularization parameters for the LC, LQC, LQPC and LQPTC runs. The optimal ranges for different feature class settings were determined by visual inspection of m-curves.

Evaluating Maxent tuning using presence-absence test data

We evaluated our tuning of Maxent's parameters by comparing with the best possible settings for the given evaluation data: those that yield the best performance when models are built from the presence-only training data and evaluated on the presence-absence test data. We call the latter parameter settings ''pa-tuned'', in contrast to the ''po-tuned'' values determined in ''Tuning regularization parameters using presence-only data'', above and ''Combining continuous and categorical variables'', above. In po-tuning, a single set of regularization parameters was determined for application across all regions. However, it is conceivable that different sets of parameters may be appropriate for different regions, and better performance may be obtained by tuning the parameters for each region separately. In pa-tuning, we therefore distinguished two cases: regional tuning and global tuning. In the former case, a separate set of regularization parameters was chosen to maximize the average AUC of the species in the relevant region only. In the latter case, a single set of regularization parameters was chosen to maximize the average AUC across all species.

Sets of pa-tuned regularization parameters were obtained by local search, a heuristic optimization technique that tries to improve the value of the objective (in our case the AUC on the presence-absence data) by making incremental changes in parameters. We began with the po-tuned parameter values and then cyclically iterated through feature classes, trying to increase or decrease the corresponding regularization parameter. This was repeated until no changes in parameters yielded an improvement. We considered multiplicative changes in regularization parameters by a factor of $\sqrt{2}$ or $1/\sqrt{2}$. We allowed at most an 8-fold increase or decrease relative to the po-tuned parameter setting. Each feature class was applied in the same range of numbers of occurrences as determined by po-tuning in ''Choosing optimal features sets'', above.

Both regional and global tuning were performed using two different choices of training background sets. We used a random sample of 10 000 sites from the study area (random background) and the set of all occurrence records for the target group (target-group background, see below).

Experiments for new extensions to Maxent

Hinge features

To evaluate the benefit of using hinge features, we determined their regularization parameter β_H and the minimum number of occurrence records for which they should be used, using presence-only data and the same methodology as in ''Experiments for tuning Maxent'', above. We then generated models for all 226 species and evaluated them using the independent presence-absence test data, comparing the resulting AUC and COR values with those obtained without hinge features.

Logistic output

Logistic, cumulative and raw outputs are all monotonically rated, i.e. they rank sites in the same order, so their statistical performance is identical when measured using AUC. On the other hand, their performance may differ when measured using COR, which depends on the actual output values rather than just their ordering. We therefore ran Maxent with all three output formats on the full suite of 226 species, and compared the three sets of predictions using the COR statistic evaluated using the independent presence/absence test data.

Sample selection bias and target-group background

Section ''Maxent modeling of distributions'' above assumes that occurrences are unbiased samples from the species' distribution, but that assumption is often not valid for occurrence data (Reddy and Dávalos 2003). A simple strategy to remove sample selection bias is to replace the uniform background data by a random sample of background data drawn from the sampling distribution. As a result, both the background data and species presences become biased in the same manner. Since Maxent is choosing the distribution of maximum entropy relative to

the provided background, the sample selection bias is effectively factored out (Dudík et al. 2005). When using such a background sample, it may happen that there is no probability distribution over the background sites that satisfies all the constraints derived from the occurrence data. Therefore, in order to ensure that the optimization problem as described in ''Maxent modeling of distributions'', above remains feasible, the species' presence sites must be added to the background sample, if they are not already there. The Maxent distribution is thus calculated over the union of the background and presence sites, yielding a Gibbs distribution which can then be projected onto the rest of the study area (or any other area). We investigated whether the set of occurrence data for an entire ''target group'' of species that may be captured or observed using the same methods (Ponder et al. 2001, Anderson 2003) can be regarded as a random sample from the sampling distribution and thus used as background data. We generated models for all 226 species using target-group background data and compared their predictive performance with models generated using random background. The regions awt and nsw contained multiple target groups: region awt contained birds and plants; region nsw contained birds, reptiles, mammals, and plants. The remaining regions had only a single target group.

Random background sampling

The runtime of Maxent scales linearly with the size of the background set. Thus, for large, fine-grain maps, Maxent may be prohibitively slow. However, because of the considerable uncertainty in the presence data, a large amount of background data may not be necessary. Instead, we can use a small subsample of the background to significantly reduce runtime without sacrificing predictive performance. Here we investigate how large a background sample is needed. We chose a geometric progression of background sizes, namely the set {63, 125, 250, 500, 1000, 2000, 4000, 8000, 16 000, 32 000, 64 000, 128 000, 256 000}. For each background size, we generated models for all 226 species for each of 10 random background sets of that size, determining the average AUC using the independent presence/absence test data. All models were generated using regularization parameters and feature classes as tuned in ''Experiments for tuning Maxent'' and ''Hinge features'', above.

Results

Tuning of Maxent's parameters

By visual inspection of β-curves for L, LQ, LQP, T, and C runs, we propose the regularization parameter settings given in the top portion of Table 3. Values set in boldface were determined exactly, while the others were obtained from the boldface values by interpolation or extrapolation. Values between the highest and lowest bold settings are linearly interpolated between the two closest bold settings. Values above the highest bold settings are kept the same as the highest bold settings, and the values below the lowest bold settings are linearly extrapolated from the two lowest bold settings. Note that feature classes with larger numbers of features generally have higher β values, consistent with the theory in ''Maxent modeling of distributions'', above. However, for each feature class, the best value of β is monotonically non-increasing in the number of occurrence records. In other words, model performance is optimized if we use error bounds that decrease in width somewhat faster than $1/\sqrt{\text{sample size}}$, as the theory suggests. The β-curves were generally smooth and unimodal (Fig. 2).

When using both continuous and categorical variables, the intermediate setting for β_C (Table 3) gave the best performance in more than half of the m-curves and never gave the worst performance. Discrete ordinal variables performed best when viewed as continuous (details not shown).

Finally, we determined optimal combinations of feature classes. Figure 3 shows the performance of four feature class settings. From the figures, we determined ranges of individual feature classes as follows: LC features for 2–9 samples, LQC features for 10–79 samples, LQPTC features for 80 and more samples.

Evaluation of presence-only tuning using presence-absence data

In Table 5, we compare performance of po-tuned Maxent parameters and pa-tuned parameters. Global pa-tuning results in an improvement of average AUC by 0.006

Table 3. Parameter settings tuned using presence-only data. Values in boldface were determined exactly, values in italics are linearly interpolated or extrapolated, with the exception that the values to the right of the listed ranges remain constant. For category indicator features, the ''low'' settings were determined using a single (categorical) variable, while the intermediate settings were chosen to approximate the geometric average of the ''low'' setting and β_L.

	Number of occcurrence records					
	0	6	10	17	30	100
Linear features: β_L	*1.0*	**1.0**	**1.0**	*0.72*	**0.2**	**0.05**
Linear and quadratic features: β_L, β_Q	*1.3*	**1.0**	**0.8**	**0.5**	**0.25**	**0.05**
Linear, quadratic and product features: β_L, β_Q, β_P	*2.6*	**2.0**	**1.6**	**0.9**	**0.55**	**0.05**
Threshold features: β_T	*2.0*	*1.94*	**1.9**	*1.83*	*1.7*	**1.0**
Category indicators, a single categorical variable, ''low'': β_C	0.2	**0.2**	**0.2**	**0.1**	**0.05**	*0.05*
Category indicators, ''intermediate'': β_C	0.65	0.53	0.45	0.25	0.15	0.05
Hinge features: β_H	**0.5**	**0.5**	**0.5**	**0.5**	**0.5**	**0.5**

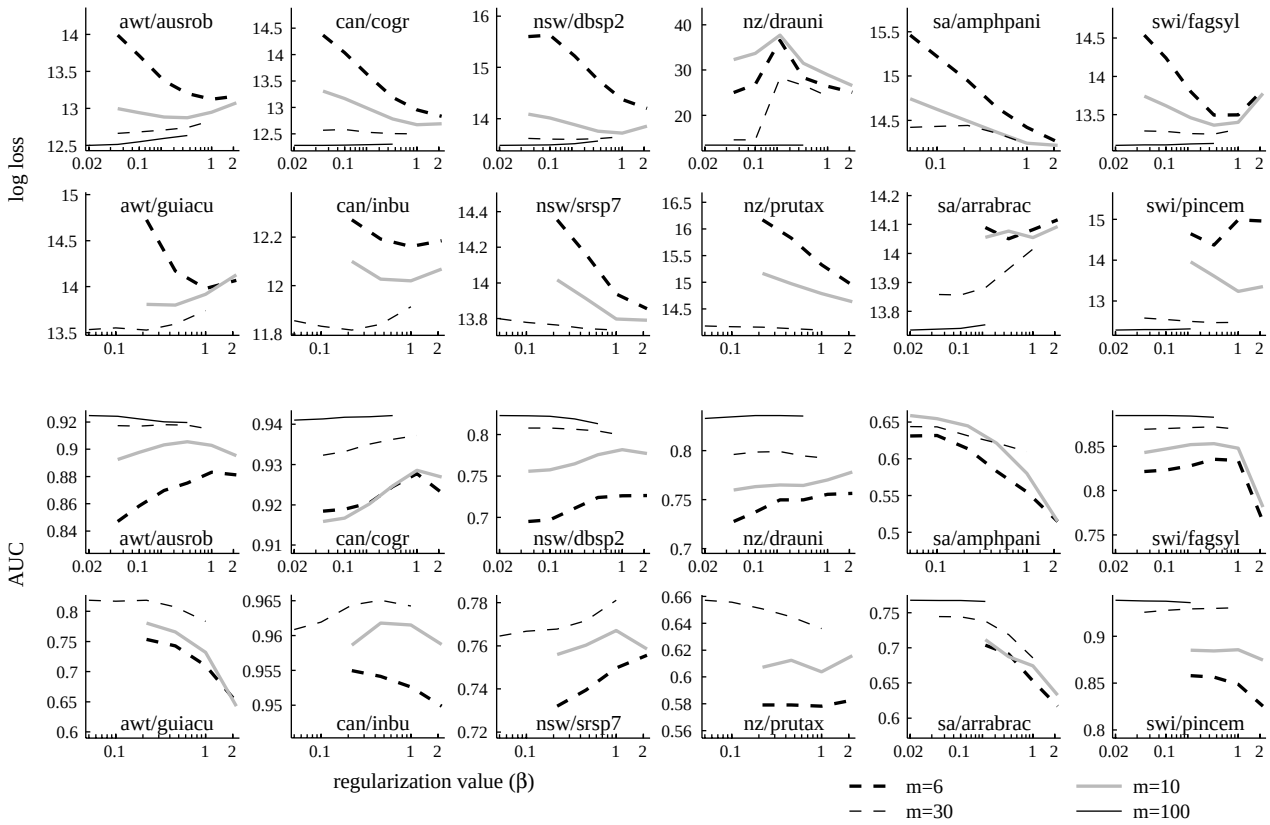

Fig. 2. A subset of the β-curves for LQ runs of Maxent. Average performance of Maxent is evaluated in terms of log loss and AUC over five random partitions as a function of the regularization parameter β (where $\beta_L = \beta_Q = \beta$) for varying number of occurrences m. Default values of β_L, β_Q for Maxent ver 1.8.3 through at least 3.1.0 (listed in Table 3) were chosen to obtain satisfactory performance on the evaluated subset of species. For explanations of species codes, see Table 2.

(both for random background and target-group background) whereas regional pa-tuning improves the AUC for random background by 0.014 and the AUC for target-group background by 0.012.

In Table 6, we report parameters obtained by global pa-tuning and medians of parameters obtained by regional pa-tuning. For regions awt and nsw, the regional tuning was performed separately for each taxonomic group (birds and plants in awt, and small mammals, reptiles, birds and plants in nsw), resulting in a total of 10 regionally optimized parameter sets. Each median is thus taken over a set of 10 values.

To compare pa-tuned regularization parameters with po-tuned regularization parameters, we determined the median training set size in each range and report the corresponding po-tuned values. Note that the pa-tuned values are almost always larger than the po-tuned values. Larger regularization represents our increased uncertainty in feature-expectation estimates as a result of differences between training and test distributions.

Table 4. Average values of the COR statistic for Maxent run with either random or target-group background, and with output given in either raw, cumulative or logistic format.

Background type	Raw	Cumulative	Logistic
Random	0.184	0.206	0.210
Target-group	0.238	0.239	0.245

New extensions to Maxent

Hinge features

Tuning on the presence-only data indicated that hinge features should be used when the number of presence records is at least 15, with a regularization parameter of 0.5. Using these parameters, a full run on the 226 species with random background yielded an average AUC of 0.726, which was significantly better than the average of 0.721 obtained without hinge features ($p < 0.001$, two-tailed Wilcoxon signed-rank test, pairing by species). Similarly significant improvement was observed for the COR statistic (from 0.198 to 0.210, $p < 0.001$).

Logistic output

When compared using the COR statistic, the logistic output format achieved the best performance, followed by cumulative output format and then raw output (Table 4), and the ordering was consistent when using either random background or target-group background. The performance of logistic output was significantly better than raw output when using random background, and significantly better than cumulative output when using target-group background ($p < 0.01$, two-tailed Wilcoxon signed-rank test, pairing by species).

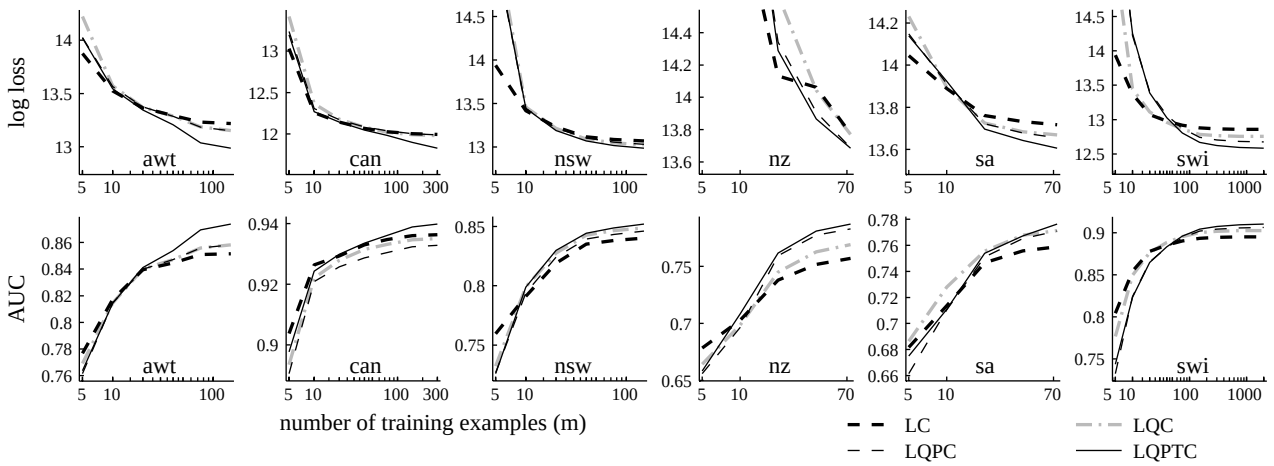

Fig. 3. Curves showing average performance of Maxent as a function of sample size (m-curves) averaged over all species selected for tuning. By visual inspection, we determined the ranges of sample sizes in which to use the different sets of feature classes as: LC features for 2–9 samples, LQC features for 10–79 samples and LQPTC features for 80 and more samples. AUC values were calculated using background data in place of true absences, as the data used to generate these curves is presence-only. Each region is shown separately; for region names, see Table 1.

Background treatments

Using target-group background instead of random background raised the average AUC from 0.726 to 0.757, as measured using independent presence-absence test data. This improvement is highly significant ($p < 10^{-6}$, one-tailed sign test, pairing by species). The results of the background sampling experiment (Fig. 4) indicate that predictive performance increases substantially as the number of background sites increases, reaching a plateau after 8000 sites.

Discussion and conclusions

Maxent tuning

A major focus of this work is the tuning of regularization parameters and choice of feature classes, both viewed as functions of the number of presence samples available during training. Tuning was performed on presence-only data (po-tuned parameters) and then compared to parameters optimized for presence-absence evaluation data (pa-tuned parameters). The po-tuning procedure was robust: five-fold cross-validation reduced the variance of parameter estimates, as did the fact that we took averages or medians over several species to determine each parameter. In addition, the po-tuned values were evaluated on many more species in the presence-absence evaluation data, and we found that the po-tuned parameters resulted in model performance which is almost as good as if the parameters had been tuned on the evaluation data itself. The datasets used here cover a wide range of species, taxonomic groups, numbers of occurrence records and species prevalence (fraction of the study area occupied by the species). We conclude therefore that the tuning methodology based on presence-only data was very effective, and the resulting regularization parameters, which are the defaults in Maxent software ver. 1.8.3 through at least 3.1.0 (with one exception, described below), are well suited for a wide range of presence-only datasets. It is tempting to use the pa-tuned settings as default settings in the software, since they give marginally better performance on the evaluation data. However, we believe that doing so may result in overfitting to this particular evaluation dataset, since the pa-tuned

Table 5. Maxent performance in terms of AUC under four different training-evaluation scenarios: presence-only training on random background and target-group background, and evaluation with respect to globally or regionally presence-absence optimized parameters. AUC values were determined using the presence-absence data.

	Random background			Target-group background		
		improvement from default			improvement from default	
	default settings	globally optimized settings	regionally optimized settings	default settings	globally optimized settings	regionally optimized settings
awt	0.693	0.004	0.015	0.729	0.000	0.009
can	0.594	0.008	0.023	0.719	0.011	0.020
nsw	0.711	0.005	0.022	0.742	0.009	0.020
nz	0.733	0.008	0.009	0.741	0.009	0.011
sa	0.796	0.007	0.014	0.793	0.003	0.005
swi	0.803	0.003	0.001	0.837	0.006	0.006
all species	0.726	0.006	0.014	0.757	0.006	0.012

Table 6. Overview of pa-tuned parameters: globally optimized parameters and medians of 10 regionally optimized parameters. The global settings optimize the average performance across all species. The regional settings are optimized separately for each of 10 taxonomic groups in the 6 regions.

		Random background				Target-group background			
		number of occurrences				number of occurrences			
		2–9	10–14	15–79	≥80	2–9	10–14	15–79	≥80
β_L:	global optimum	1.00	1.41			2.00	1.41		
	regional median	1.00	1.00			1.00	1.00		
	default*	1.00	0.71			1.00	0.71		
β_Q:	global optimum		1.41	0.50	0.35†		2.00	1.00	0.35†
	regional median		1.00	0.85	0.05		1.00	0.71	0.07
	default*		0.71	0.23	0.05		0.71	0.23	0.05
β_P:	global optimum				0.35†				0.35†
	regional median				0.04				0.11
	default*				0.05				0.05
β_T:	global optimum				2.00				8.00†
	regional median				1.21				1.71
	default*				1.00				1.00
β_H:	global optimum			0.35	0.50			0.71	1.41
	regional median			0.85	0.50			1.21	0.50
	default*			0.50	0.50			0.50	0.50
β_C:	global optimum	1.41	0.50	0.03**	0.03	0.50	1.00	0.13	0.25
	regional median	0.71	0.50	0.18	0.04	0.71	0.50	1.00	0.04
	default*	0.53	0.39	0.14	0.05	0.53	0.39	0.14	0.05

* po-tuned values are listed for the median training set size appearing in each range: 6, 12, 36, and 221.
† the largest possible value in local search.
** the smallest possible value in local search.

settings are being evaluated here on the same data on which they were tuned. We believe it is preferable to continue using the po-tuned settings, which have been validated on independent test data in this study.

We emphasize, though, that datasets that deviate significantly from those used in this study may require further parameter tuning (for example, using the tuning approach demonstrated here) to optimize Maxent model performance. The six regional datasets we used all contain a similar number of environmental variables (11–13). The theory in "Maxent modeling of distributions", above suggests that more regularization may be needed if the number of environmental variables is much larger. Similarly, $>10\,000$ background sites may be needed if the number of presence sites is much greater than for the species studied here. In addition, the datasets used here made little use of categorical data, and since the time when we performed this study, some other datasets with more categorical data have shown strong signs of overfitting with the regularization as tuned here. For this reason, software ver. 2.3.35 and later use higher categorical regularization, with the second last line of Table 3 being increased to {0.65, 0.56, 0.5, 0.25, 0.25, 0.25}. This change has minimal effect on Maxent's statistical performance on the data we have used here.

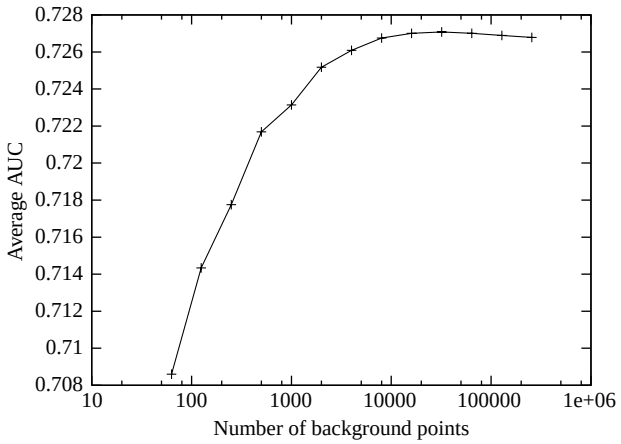

Fig. 4. Average performance of Maxent for varying background sample sizes. For each sample size, 10 random background samples were taken, and for each such sample, models were generated for all 226 species. Values shown are average AUC values across all random samples and species. The AUC values were calculated using independent presence-absence data.

Possible reasons for Maxent's good performance

The parameter settings developed in this study were used (without hinge features or logistic output, which came later) in the modeling comparison of Elith et al. (2006), and we feel that our careful parameter tuning, especially for small sample sizes, contributed to Maxent's good showing in that comparison. There are a number of other factors that may help explain Maxent's good performance.

First, Maxent uses ℓ_1 regularization, which tends to produce models with few non-zero coefficients (Williams 1995, Tibshirani 1996) and therefore encourages parsimony. Regularization appears to prevent overfitting better than variable-selection methods commonly used for regression-based models such as generalized additive and generalized linear models. These models can also use regularization – ℓ_1 regularization applied to such methods is known as the lasso – but the regularized variants have not been used for species distribution modeling to date.

Perhaps more importantly, when used on presence-only data, regression-based methods suffer from the problem of "contaminated controls" (Keating and Cherry 2004), in

Table 7. Average performance for current versions and settings of Maxent and boosted regression trees (BRT), evaluated on independent presence/absence test data using area under the receiver operating characteristic curve (AUC), correlation with 1/0 values for presence/absence (COR), and average deviance per test site. Each value shown is the average over 226 species.

Algorithm	Background type	AUC	COR	Average deviance
Maxent	Random	0.7276	0.2100	0.9929
Maxent	Target-group	0.7569	0.2446	0.8585
BRT	Random	0.7275	0.2130	1.1249
BRT	Target-group	0.7544	0.2435	0.9635

which background data is treated as absence data, even though it is contaminated with presences. Recent work addresses this issue by using the EM algorithm to essentially decontaminate the background (Ward et al. 2007), but this approach has not been extensively compared to other presence-only modeling methods because it requires knowledge of the species prevalence in the study region, which cannot be inferred from presence-only data. A final possible explanation for Maxent's good performance is that it is a generative approach, modeling the species distribution $P(x|y=1)$ directly, whereas regression-based methods are discriminative, modeling $P(y=1|x)$. Maxent's good performance on small samples is in line with previous studies indicating that generative methods give better predictions than discriminative methods when the amount of training data is small (Ng and Jordan 2001).

Lastly, we note that boosted regression trees (Friedman 2001) slightly out-performed Maxent in the comparison of Elith et al. (2006). Both methods have been further developed since the comparison: for BRT, there is improved tuning for presence-only data (J. Elith pers. comm.), and for Maxent, we introduced hinge features and logistic output format. The current versions of BRT and Maxent achieve average AUC and COR scores that are extremely close; however, Maxent achieves somewhat better average deviance scores (Table 7). This discrepancy is probably because Maxent's logistic format is robust to the unknown prevalence (''Maxent output formats and logistic models'', above), while BRT (as applied to presence-only data) gives equal weight to presences and absences (Elith et al. 2006), which seems to inflate estimates of probability of presence.

Maxent extensions

We have shown that target-group background can significantly improve model performance. The magnitude of the increase in AUC that results from using target-group background is similar to the differences in AUC between modeling methods in Elith et al. (2006). The performance improvement was greatest in Ontario, Canada, the region with the most glaring sample selection bias in the presence-only training data. Target-group background therefore appears effective at countering sample selection bias, matching the theoretical prediction (Dudík et al. 2005) and suggesting that in large natural history museum and herbarium collections, target-group background may be interpreted as a random sample from the (biased) sampling distribution. Target-group background can also be used for other modeling methods, and is analyzed in detail elsewhere (Phillips et al. unpubl.). We have also shown that using a random sample of around 10 000 background sites achieves the same model performance as using the full set of sites from the whole study area, offering a dramatic reduction in processing requirements for large datasets.

The new hinge features introduced here significantly improve model performance. Hinge features can be used with any presence-only dataset, and are used by default in the software whenever there are at least 15 presence sites. In fact, hinge features can effectively replace quadratic, product and threshold features: when hinge features are used, omitting Q, P and T features hardly changes predictive performance, with average AUC and COR scores changing by <0.0001 for both random and target-group background. Interestingly, hinge features do not significantly increase the complexity of models that Maxent can produce, since threshold features already allow an arbitrary response to each environmental variable. However, hinge features make piecewise linear contributions to the exponent of the Maxent model, which appear to incur less regularization penalty when approximating species' true response to the environmental variables than the step-function response produced by threshold features. In other words, hinge features allow simpler and more succinct approximations of the true species response to the environment. Building on this principle, an interesting topic for future research would be to develop new feature classes that allow even more succinct response approximations, such as the splines commonly used in generalized additive models (Hastie and Tibshirani 1990).

The logistic output format, introduced here, is easier to interpret than previous Maxent output formats: it can be interpreted as estimating a species probability of presence, conditioned on environmental variables. For sites with small logistic value, i.e. that are predicted to be unsuitable or marginally suitable for the species, the logistic value is proportional to Maxent's raw output. For the most suitable sites, the logistic value is capped so that it remains below one. This addresses a problem with raw output, that the exponential model is not bounded above and can therefore give unreasonably high values to some sites, corresponding to probability of presence greater than one. Because the logistic format fixes this problem, it is better calibrated, in the sense that it has improved average COR.

Beyond the realized distribution

The evaluation and tuning we have described measures model performance according to the ability to predict the realized distribution of a species, and the parameter tunings are therefore optimized for predicting realized distributions. It is important to note that many applications of species distribution models depend on predicting potential distributions, rather than realized distributions. A species may have failed to disperse due to geographic barriers, or be excluded from an area due to competition. In the current evaluation, models which predict into such areas would be penalized, though it is not clear how many of the species considered here are absent from significant portions of their

potential distribution. Some applications, however, require prediction into unoccupied areas, for instance, when measuring the inter-predictivity of models of sister species (Peterson et al. 1999). In such cases, good performance on presence-absence data from the realized distribution may not necessarily imply effective prediction of the potential distribution. Thus, it is possible that different parameter settings would be needed than those developed here. When attempting to predict potential distributions, care should be taken to avoid indirect predictor variables whose effect on species may vary across the modeled region, such as altitude or the climatic conditions in particular months of the year (Peterson 2006, Phillips et al. 2006). Care is also needed with the choice of background data used during training, for example by removing areas where the species is excluded by known geographic barriers.

Other uses of species distribution models involve ''transferral'': producing a model over one study area and then applying it to another area, or to changed environmental conditions in the same area. Example applications include predicting the effect of climate change on species distributions (Thomas et al. 2004, Araujo et al. 2005) and predicting areas at risk for species invasions (Peterson et al. 2003, Thuiller et al. 2005). Such applications may require different choices of feature types and regularization parameters from those defined here. In particular, it has been shown that modeling methods that otherwise produce reasonably similar predictions can make wildly different predictions of species range size under climate change (Pearson et al. 2006). It is an important avenue of future research to determine guidelines for using Maxent (and other modeling methods) to create models that can reliably be transferred to alternate climatic conditions or geographic areas.

Acknowledgements – We thank the other members of the working group on ''Testing alternative methodologies for modelling species' ecological niches and predicting geographical distributions'', led by A. T. Peterson and C. Moritz and hosted by the National Center for Ecological Analysis and Synthesis in Santa Barbara. Their kind inclusion of Maxent in the study enabled much of the research described here. We further thank those who supplied data for the working group, including working group members as well as A. Ford, CSIRO Atherton, for AWT data; M. Peck and G. Peck, Royal Ontario Museum and M. Cadman, Bird Studies Canada, Canadian Wildlife Service of Environment Canada, for CAN data; the National Vegetation Survey Databank and the Allan Herbarium, for NZ data; Missouri Botanical Garden, especially R. Magill and T. Consiglio, for SA data; and T. Wohlgemuth and U. Braendi from WSL Switzerland for SWI data. We also wish to thank Rob Anderson and Catherine Graham for extremely helpful commentary on the manuscript. Miroslav Dudík received support through NSF grant CCR-0325463.

References

Anderson, R. P. 2003. Real vs. artefactual absences in species distributions: tests for *Oryzomys albigularis* (Rodentia: Muridae) in Venezuela. – J. Biogeogr. 30: 591–605.

Araújo, M. B. et al. 2005. Validation of species-climate impact models under climate change. – Global Change Biol. 11: 1504–1513.

Cover, T. M. and Thomas, J. A. 2006. Elements of information theory, 2nd ed. – Wiley.

Dudík, M. et al. 2004. Performance guarantees for regularized maximum entropy density estimation. – In: Proceedings of the Seventeenth Annual Conference on Computational Learning Theory. ACM Press, New York, pp. 655–662.

Dudík, M. et al. 2005. Correcting sample selection bias in maximum entropy density estimation. – In: Advances in Neural Information Processing Systems 18. The MIT Press, pp. 323–330.

Elith, J. 2002. Quantitative methods for modeling species habitat: comparative performance and an application to Australian plants. – In: Ferson, S. and Burgman, M. (eds), Quantitative methods for conservation biology. Springer, pp. 39–58.

Elith, J. and Leathwick, J. 2007. Predicting species distributions from museum and herbarium records using multiresponse models fitted with multivariate adaptive regression splines. – Divers. Distrib. 13: 265–275.

Elith, J. et al. 2006. Novel methods improve prediction of species' distributions from occurrence data. – Ecography 29: 129–151.

Fielding, A. H. and Bell, J. F. 1997. A review of methods for the assessment of prediction errors in conservation presence/absence models. – Environ. Conserv. 24: 38–49.

Friedman, J. 1991. Multivariate adaptive regression splines (with discussion). – Ann. Stat. 19: 1–141.

Friedman, J. H. 2001. Greedy function approximation: a gradient boosting machine. – Ann. Stat. 29: 1189–1232.

Graham, C. H. et al. 2004. New developments in museum-based informatics and applications in biodiversity analysis. – Trends Ecol. Evol. 19: 497–503.

Graham, C. H. et al. 2006. Habitat history improves prediction of biodiversity in a rainforest fauna. – Proc. Nat. Acad. Sci. USA 103: 632–636.

Grünwald, P. 2000. Maximum entropy and the glasses you are looking through. – In: Proceedings of the Sixteenth Annual Conference on Uncertainty in Artificial Intelligence (UAI2000), pp. 238-246.

Grünwald, P. D. and Dawid, A. P. 2004. Game theory, maximum entropy, minimum discrepancy, and robust Bayesian decision theory. – Ann. Stat. 32: 1367–1433.

Hastie, T. and Tibshirani, R. 1990. Generalized additive models. – Chapman and Hall.

Hastie, T. et al. 2001. The elements of statistical learning: data mining, inference, and prediction. – Springer.

Jaynes, E. T. 1957. Information theory and statistical mechanics. – Phys. Rev. 106: 620–630.

Keating, K. A. and Cherry, S. 2004. Use and interpretation of logistic regression in habitat-selection studies. – J. Wildl. Manage. 68: 774–789.

Ng, A. Y. and Jordan, M. I. 2001. On discriminative vs. generative classifiers: a comparison of logistic regression and naive Bayes. – Adv. Neural Inform. Process. Syst. 14: 605–610.

Pearson, R. G. et al. 2006. Model-based uncertainty in species range prediction. – J. Biogeogr. 33: 1704–1711.

Peterson, A. T. 2006. Uses and requirements of ecological niche models and related distributional models. – Biodiv. Inform. 3: 59–72.

Peterson, A. T. et al. 1999. Conservatism of ecological niches in evolutionary time. – Science 285: 1265–1267.

Peterson, A. T. et al. 2003. Predicting the potential invasive distributions of four alien plant species in North America. – Weed Sci. 51: 863–868.

Phillips, S. J. et al. 2004. A maximum entropy approach to species distribution modeling. – In: Proceedings of the Twenty-First International Conference on Machine Learning. ACM Press, New York, pp. 472–486.
Phillips, S. J. et al. 2005. Maxent software for species distribution modeling. – <http://www.cs.princeton.edu/~schapire/maxent/>.
Phillips, S. J. et al. 2006. Maximum entropy modeling of species geographic distributions. – Ecol. Model. 190: 231–259.
Ponder, W. F. et al. 2001. Evaluation of museum collection data for use in biodiversity assessment. – Conserv. Biol. 15: 648–657.
Randin, C. F. et al. 2006. Are niche-based species distribution models transferable in space? – J. Biogeogr. 33: 1689–1703.
Reddy, S. and Dávalos, L. M. 2003. Geographical sampling bias and its implications for conservation priorities in Africa. – J. Biogeogr. 30: 1719–1727.
Thomas, C. D. et al. 2004. Extinction risk from climate change. – Nature 427: 145–148.
Thuiller, W. et al. 2005. Niche-based modelling as a tool for predicting the risk of alien plant invasions at a global scale. – Global Change Biol. 11: 2234–2250.
Tibshirani, R. 1996. Bias, variance and prediction error for classification rules. – Technical report, Univ. of Toronto.
Topsøe, F. 1979. Information theoretical optimization techniques. – Kybernetika 15: 8–27.
Ward, G. et al. 2007. Presence-only data and the em algorithm. – Biometrics, in press.
Williams, P. M. 1995. Bayesian regularization and pruning using a Laplace prior. – Neural Comput. 7: 117–143.
Zaniewski, A. E. et al. 2002. Predicting species spatial distributions using presence-only data: a case study of native New Zealand ferns. – Ecol. Model. 157: 261–280.
Zheng, B. and Agresti, A. 2000. Summarizing the predictive power of a generalized linear model. – Stat. Med. 19: 1771–1781.